CS 7545: Machine Learning Theory **Fall 2021**

Lecture 14

Name: Tyler LaBonte **Professor:** Santosh S. Vempala

1 PAC-learning and Uniform Convergence

We continue with the proofs of the theorems from the previous lecture. Recall the Chernoff bound: if $X_1, X_2, \ldots, X_m$ are independent Bernoulli random variables and $X = \sum_{i=1}^m X_i$, then

$$\Pr[X \geq (1+\delta)\mathbb{E}[X]] \leq \exp\left(-\frac{\delta^2}{2+\delta}\mathbb{E}[X]\right) \tag{1}$$

$$\Pr[X \geq (1-\delta)\mathbb{E}[X]] \leq \exp\left(-\frac{\delta^2}{2}\mathbb{E}[X]\right). \tag{2}$$

Also useful is the Hoeffding bound: if the X_i are as above with $a_i \leq X_i \leq b_i$, then

$$\Pr[X \geq \mathbb{E}[X] + t] \leq \exp\left(\frac{-t^2}{\sum_{i=1}^m (b_i - a_i)^2}\right) \tag{3}$$

$$\Pr[X \leq \mathbb{E}[X] + t] \leq \exp\left(\frac{-t^2}{\sum_{i=1}^m (b_i - a_i)^2}\right). \tag{4}$$

Proof: (of the PAC-learning theorem). We need to show that $\text{err}_S(h) = 0$ implies $\text{err}_D(h) \leq \epsilon$ with probability at least $1 - \delta$. The idea of a "bad sample" which can cause us to fail this condition is event A: $\text{err}_S(h) = 0$ and $\text{err}_D(h) > \epsilon$. We will show that $\Pr[A] \leq \delta$.

Consider two subsets S, S' of size m. Let B be the event that $\text{err}_S(h) = 0$ but $\text{err}_{S'}(h) \geq \frac{\epsilon}{2}$.

Claim 1.1

$\Pr[B] \geq \frac{1}{2}\Pr[A]$. (This enables us to work with two finite sets rather than a set and the entire distribution).

Proof: Clearly $\Pr[B] \geq \Pr[A]\Pr[B|A]$. The latter is the event that h has error at least $\frac{\epsilon}{2}$ on m points drawn from D given that it has error at least ϵ on D. Let X_i be the indicator random variable for h having an error on point i. Then by the Chernoff bound,

$$\Pr\left[\sum_i X_i < \mathbb{E}\left[\sum_i X_i\right] - \frac{\epsilon m}{2}\right] \leq \exp\left(-\frac{\epsilon m}{8}\right). \tag{5}$$

Thus, if $m \geq \frac{8}{\epsilon}$, $\Pr[B|A] \geq \frac{1}{2}$. ■

Thus, it suffices to show $\Pr[B] \leq \frac{\delta}{2}$. We will choose S and S' in a specific way to help us analyze B. Suppose we draw a set S'' of size $2m$ from D and partition it into S and S' in the following

manner: we pair up points as $(a_1, b_1), (a_2, b_2), \ldots, (a_m, b_m)$, and from each pair, one element goes to S and one element goes to S'.

Clearly (a_i, b_i) cannot both be errors, because then S would have an error, and we assumed $\text{err}_S(h) = 0$. On the other hand, the number of pairs for which there is one error is at least $\frac{\epsilon m}{2}$, since we assumed $\text{err}_{S'}(h) > \frac{\epsilon}{2}$. Thus, all these error points need to be assigned to S', which happens with probability

$$\Pr[B] \leq 2^{-\frac{\epsilon m}{2}}. \tag{6}$$

This shows $\Pr[B]$ for a fixed hypothesis. For all hypotheses,

$$2^{-\frac{\epsilon m}{2}} \mathcal{H}[2m] \leq \frac{\delta}{2} \implies m \geq \frac{2}{\epsilon}\left(\log 2\mathcal{H}[2m] + \log\frac{1}{\delta}\right). \tag{7}$$

■

Proof: (of the uniform convergence theorem). We can adapt the previous proof. Let A be the event that $|\text{err}_S(h) - \text{err}_D(h)| \geq \epsilon$ and B be the event that $\text{err}_S(h) - \text{err}_{S'}(h)| \geq \frac{\epsilon}{2}$. We want $\Pr[B] \leq \frac{\delta}{2}$ to show $\Pr[A] \leq \delta$.

We can select S'' and partition it as above. We don't care about cases where both points in a pair are correct or wrong, since it doesn't affect the relative error. We know there are at least $\frac{\epsilon m}{2}$ steps with an error, and we want to bound how many errors can be randomly partitioned into a certain set. Let X_i be 1 if the error goes to S and -1 if the error goes to S'. Then by the Hoeffding bound,

$$\Pr\left[\left|\sum_i X_i\right| \geq \frac{\epsilon m}{2}\right] \leq 2\exp\left(\frac{-\epsilon^2 m}{8}\right). \tag{8}$$

Hence,

$$2\exp\left(\frac{-\epsilon^2 m}{2}\right)\mathcal{H}[2m] \leq \frac{\delta}{2} \implies m \geq \frac{8}{\epsilon^2}\left(\log 4\mathcal{H}[2m] + \log\frac{1}{\delta}\right). \tag{9}$$

■

The major difference between the two theorems is the ϵ vs ϵ^2 dependence. This is essentially because an additive error of ϵ is much easier to obtain when the expectation is around zero than when the expectation is large. The fundamental reason for this can be understood by studying the proofs of the concentration bounds.